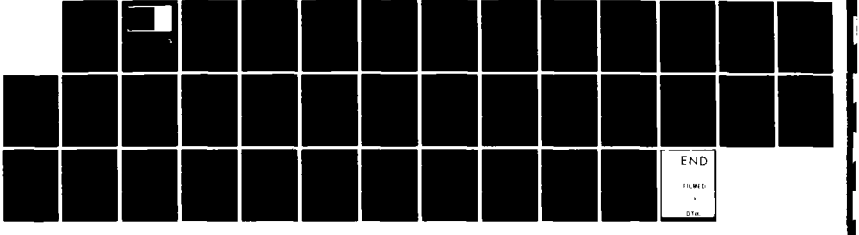
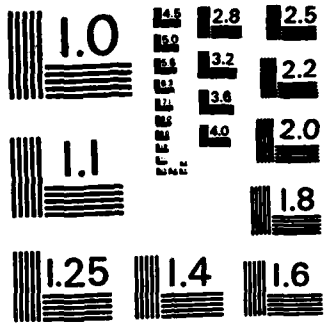
AD-A120 956 PREDICTION AND ENTROPY(U) WISCONSIN UNIV-MADISON 1/1
MATHEMATICS RESEARCH CENTER H AKAIKE JUN 82
MRC-TSR-2397 DAAG29-80-C-0041

UNCLASSIFIED F/G 20/13 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

MRC Technical Summary Report #2397

PREDICTION AND ENTROPY

Hirotugu Akaike

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

June 1982

(Received May 4, 1982)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

PREDICTION AND ENTROPY

Hirotugu Akaike*

Technical Summary Report #2397
June 1982

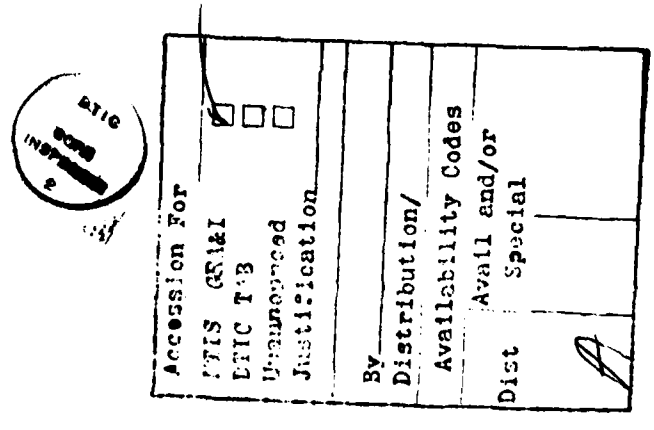
ABSTRACT

The emergence of the magic number 2 in recent statistical literature is explained by adopting the predictive point of view of statistics with entropy as the basic criterion of the goodness of a fitted model. The historical development of the concept of entropy is reviewed and its relation to statistics is explained by examples. The importance of the entropy maximization principle as the basis of the unification of conventional and Bayesian statistics is discussed.

AMS (MOS) Subject Classifications: 62-02, 62A99, 62F99

Key Words: Entropy, Predictive distribution, Likelihood, Model selection, Bayes procedure, Information, AIC, Entropy maximization principle

Work Unit Number 4 (Statistics and Probability)



*Address: The Institute of Statistical Mathematics, Tokyo, Japan.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

This work is concerned with the clarification of the importance of the roles played by the predictive point of view and the concept of entropy in statistics. It starts with the discussion of the appearance of a common constant, the so called magic number 2, in various applications of statistics, and shows that it is deeply related to the predictive use of statistics.

The historical development of the statistical concept of entropy by L. Boltzmann is reviewed and the common confusion caused by the adoption of the Shannon entropy is eliminated. The close relation between statistics and the Boltzmann entropy is illustrated by examples.

The objectivity of the log likelihood as the criterion of fit is explained with the aid of the entropy. The generality of the magic number 2 is then demonstrated in its relation to an information criterion AIC which is realized by combining the predictive point of view and the concept of entropy.

The discussion leads to the entropy maximization principle which specifies the object of statistics as the maximization of the expected entropy of a fitted predictive distribution. It is shown that this principle provides a basis for the unification of conventional and Bayesian statistics. Obviously the recognition of such possibility contributes significantly to the enhancement of research activity in the general area of statistics.

PREDICTION AND ENTROPY

Hirotugu Akaike*

1. Introduction and summary

In this paper we start with an observation that the emergence of a particular constant, the magic number 2, in several statistical papers is inherently related with the predictive use of statistics. The generality of the constant can only be appreciated when we adopt the statistical concept of entropy, originally developed by a physicist L. Boltzmann, as the criterion to measure the deviation of a distribution from another.

A historical review of Boltzmann's work on entropy is given to provide a basis for the interpretation of the statistical entropy. The neg-entropy, or the negative of the entropy, is often equated to the amount of information. This review clarifies the limitation of Shannon's definition of the entropy of a probability distribution. The relation between the Boltzmann entropy and the asymptotic theory of statistics is discussed briefly.

The concept of entropy provides a proof of the objectivity of the log likelihood as a measure of the goodness of a statistical model. It is shown that this observation, combined with the predictive point of view, provides a simple explanation of the generality of the magic number 2. This is done through the explanation of the AIC statistic introduced by the present author. The use of AIC is illustrated by its application to multidimensional contingency table analysis.

The discussion of AIC naturally leads to the entropy maximization principle which specifies the object of statistics as the maximization of the expected entropy of a true distribution with respect to the fitted predictive distribution. The generality of this principle is demonstrated through its application to Bayesian statistics. The necessity of

---

*Address: The Institute of Statistical Mathematics, Tokyo, Japan.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

Bayesian modeling is discussed and its similarity to the construction of the statistical model of thermodynamics by Boltzmann is pointed out. The principle provides a basis for the unification of the Bayesian and conventional statistics. Referring to Boltzmann's fundamental contribution to statistics, the paper concludes by emphasizing the importance of the research on real problems for the development of statistics.

2. Emergence of the magic number 2

Around the year of 1970 a curious emergence of a constant has been observed in a series of papers. This is the emergence of what Stone (1977a) symbolically calls the magic number 2.

The number appears in Mallow's C_p statistic for the selection of independent variables in multiple regression which is by definition

$$C_p = \frac{1}{s^2} RSS_p - n + 2p ,$$

where RSS_p denotes the residual sum of squares after regression on p independent variables, n the sample size and s^2 an estimate of the common variance σ^2 of the error terms (Mallows, 1973). The final prediction error (FPE) introduced by Akaike (1969, 1970) for the determination of the order of an autoregression is an estimate of the mean squared error of the one-step prediction when the fitted model is used for prediction. It satisfies asymptotically the relation

$$n \log FPE = n \log S_p + 2p ,$$

where n denotes the length of the time series, S_p the maximum likelihood estimate of the innovation variance obtained by fitting the p^{th} order autoregression. Both Leonard and Ord (1976) and Stone (1977a) noticed the emergence of the number as the asymptotic critical level of F-tests when the number of observations is increased.

An explanation of this number 2 can easily be given for the case of the multiple regression analysis. The effect of regression is usually evaluated by the value of RSS_p. A smaller RSS_p may be obtained by increasing the number of independent variables p. However, we know that after adding a certain number of independent variables further addition of variables often merely increases the expected variability of the estimate. When the increase of the expected variability is measured in terms of the mean squared prediction error, it will be seen that the increase is exactly equal to the expected amount of decrease of the sample residual variance RSS_p/n. Thus to convert RSS_p into an

unbiased estimate of the mean squared error of prediction we must apply <u>twice</u> the correction that is required to convert RSS_p into an unbiased estimate of $n\sigma^2$.

The appearance of the critical value 2 for the F-test discussed by Leonard and Ord (1976) is more instructive. The F-test is considered as a preliminary test of significance in the estimation of the one-way ANOVA model where K independent observations y_{jk} $(k = 1,2,\ldots,K)$ are taken from each group j $(j = 1,2,\ldots,J)$. Under the assumption that y_{jk} are distributed as normal with mean θ_j and variance σ_W^2 the F-statistic for testing the hypothesis $\theta_1 = \theta_2 = \cdots = \theta_J$ is given by

$$F = \frac{(J - 1)^{-1}S_B^2}{(J(K - 1))^{-1}S_W^2},$$

where $S_B^2 = K\Sigma_j(y_{j\cdot} - y_{\cdot\cdot})^2$ and $S_W^2 = \Sigma_j\Sigma_k(y_{jk} - y_{j\cdot})^2$ and where $y_{j\cdot}$ and $y_{\cdot\cdot}$ denote the mean of the j^{th} group and the grand mean, respectively. The final estimate of θ_j is defined by

$$\tilde{\theta}_j = \begin{cases} y_{j\cdot} & \text{if the hypothesis is rejected} \\ y_{\cdot\cdot} & \text{otherwise .} \end{cases}$$

Consider the loss function $L(\tilde{\theta},\theta) = \Sigma(\tilde{\theta}_j - \theta_j)^2$. For the simpler estimates defined by $\tilde{\theta}_j = y_{\cdot\cdot}$ and $\tilde{\theta}_j = y_{j\cdot}$ it can easily be shown that the difference of the risks of these estimates has one and the same sign as that of $E(J(K - 1))^{-1}S_W^2(F - 2)$. Thus when the sample size K is sufficiently large the choice of the critical value 2 for the F-test to select $\tilde{\theta}_j$ is appropriate.

The characteristic that is common to these papers is that the authors considered some predictive use of the models. An early example of the use of the concept of future observation to clarify the structure of an inference procedure is Fisher (1935, p. 393). The concept is explicitly adopted as the philosophical motivation in a work by Guttman (1967). In the present paper, the point of view that considers the purpose of statistics as the realization of appropriate predictions will generally be called the predictive point of view.

In the above example of the ANOVA model, if the number of groups J is increased indefinitely, the test statistic F converges to 1 under the null hypothesis. Thus the critical value of the F-test for any fixed level of significance must also converge to 1 instead of 2. As is observed by Leonard and Ord this dramatically demonstrates the difference between the conventional approach to model selection by testing with a fixed level of significance and the predictive approach. Since there is no generally accepted criterion for the selection of the level of significance the present result must be considered as a warning against the conventional testing procedure. Thus the emergence of the magic number 2 must be considered as a sign of the impending change of the paradigm of statistics. However, to fully appreciate the generality of the number, we have first to expand our view of the statistical estimation procedure.

3. From point to distribution

The risk functions considered in the preceding section were the mean squared errors of the predictions. Such a choice of the criterion is conventional but quite arbitrary. The weakness of the ad hoc definition becomes apparent when we try to extend the concept to multivariate problems.

A typical example of multivariate analysis is factor analysis. At first sight it is not at all clear how the analysis is related to prediction. In 1971, trying to extend the concept of FPE to solve the problem of determination of the number of factors, the present author came to the recognition that in factor analysis our prediction was realized through the specification of a distribution (Akaike, 1981). This observation quickly led to the observation that almost all the important statistical procedures are concerned with the realization of predictions through the specification of some distributions.

Stigler (1975) noticed the shift of the interest of statisticians from point to distribution estimation towards the end of the 19th century. However, it seems that Fisher's very effective use of the concept of parameter drew the attention of statisticians back to the estimation of a point in a parameter space. We are now in a position to return to distributions and here the basic problem is the introduction of a natural topology in the space of distributions. The probabilistic interpretation of thermodynamic entropy developed by Boltzmann provides a historically most successful example of a solution to this problem.

4. Entropy and information

The statistical interpretation of the thermodynamic entropy, a measure of the unavailable energy within a thermodynamic system, was developed in a series of papers by L. Boltzmann in 1870's. His first contribution was the observation of the monotone decreasing behavior in time of a quantity defined by

$$E = \int_0^\infty f(x,t)\log\left[\frac{f(x,t)}{\sqrt{x}}\right]dx ,$$

where $f(x,t)$ denotes the frequency distribution of the number of molecules with energy between x and $x + dx$ at time t (Boltzmann, 1872). Boltzmann showed that for a closed system, under proper assumptions of the collision process of the molecules, the quantity E can only decrease. When the distribution f is defined with the velocities and positions of the molecules the above quantity takes the form

$$E = \iint f \log f \, dx d\xi ,$$

where x and ξ denote the vectors of the position and velocity, respectively. Boltzmann showed that for some gases this quantity, multiplied by a negative constant, was identical to the thermodynamic entropy.

The negative of the above quantity was adopted by C. E. Shannon as the definition of the entropy of a probability distribution

$$H = -\int p(x)\log p(x)dx ,$$

where $p(x)$ denotes the probability density with respect to the measure dx (Shannon and Weaver, 1949).

Almost uncountably many papers and books have been written about the use of the Shannon entropy, where the quantity H is simply referred to as a measure of information, or uncertainty, or randomness. One departure from this definition of entropy is known as the Kullback-Leibler information (Kullback an Leibler, 1951) which is defined by

$$I(q;p) = \int q(x)\log\left(\frac{q(x)}{p(x)}\right)dx$$

and relates the distribution $q(x)$ to another distribution $p(x)$. Kullback (1959, p. 6)

called this quantity the mean information per observation from $q(x)$ for discrimination in favor of $q(x)$ against $p(x)$.

Much interest has been shown in the use of these quantities as measures of statistical information. However, it seems that the potential of these quantities as statistical concepts has not been fully evaluated. Apparently this is due to the neglect of Boltzmann's original work on the probabilistic interpretation of thermodynamic entropy. Karl Pearson (1929, p. 205) cites the words of D. F. Gregory "... we sacrifice many of the advantages and more of the pleasure of studying any science by omitting all reference to the history of its progress." It seems that this has been precisely the case with the development of the statistical concept of entropy or information.

5. Distribution and entropy

The work of Boltzmann (1872) produced a demonstration of the second law of thermodynamics, the irreversible increase of entropy in an isolated closed system. In answering the criticism that the proof of irreversibility is based on the assumption of a reversible mechanical process Boltzmann (1877a) pointed out the necessity of probabilistic interpretation of the result.

At that time Meyer, a physicist, produced a derivation of the Maxwell distribution of the kinetic energy among gas molecules at equilibrium as the "most probable" distribution. Pointing out the error of Meyer's proof Boltzmann (1877b) established the now well-known identity

entropy = log probability of a statistical distribution.

His reasoning was based on the asymptotic equality

$$\log \frac{n!}{n_0!n_1!\cdots n_p!} = -n \sum_{i=0}^{p} \frac{n_i}{n} \log \frac{n_i}{n}, \qquad (1)$$

where n_i denotes the frequency of the molecules at the i^{th} energy level and $n = n_0 + n_1 + \cdots + n_p$. If we put $p_i = n_i/n$ then the right hand side is equal to $nH(p)$, where

$$H(p) = -\sum_{i=0}^{p} p_i \log p_i$$

which is the Shannon entropy of the distribution $p = (p_0, p_1, \ldots, p_p)$.

Following the idea that the frequency distribution f of molecules at a thermal equilibrium is the distribution which is the most probable under the assumption of a given total energy, Boltzmann maximized

$$H(f) = -\int_0^{\infty} f \log f dx$$

under the constraints

$$\int_0^\infty f dx = N \quad \text{and} \quad \int_0^\infty x f(x) dx = L ,$$

where x denotes the energy level, N the total number of molecules and L the total energy. The maximization produces as the energy distribution $f(x) = C \exp(-hx)$ with a proper positive constant h. Boltzmann discussed in great detail that this result could be physically meaningful only for a proper definition of the energy level x, a point commonly ignored by later users of the Shannon entropy. Incidentally we notice here an early derivation of the exponential family of distributions by the constrained maximization of $H(f)$, a technique of probability distribution generation later called by the name of maximum entropy method (Jaynes, 1957).

The change of the Boltzmann's view of the energy distribution between 1872 and 1877 is quite significant. In the 1872 paper the distribution $f(x,t)$ represented a unique entity. In the 1877b paper the distribution was considered as a random sample and its probability of occurrence was the *main subject*.

Boltzmann (1878) further extended the discussion of this point. Since the probability of getting a sample frequency distribution $(w_0, w_1, \ldots, w_p)$ from a probability distribution $(f_0, f_1, \ldots, f_p)$ is given by

$$\Omega = f_0^{w_0} f_1^{w_1} \cdots f_p^{w_p} \cdot \frac{n!}{w_0! w_1! \cdots w_p!} ,$$

Boltzmann obtained an asymptotic equality

$$\ell\Omega = w_0 \ell f_0 + w_1 \ell f_1 + \cdots + w_p \ell f_p - w_0 \ell w_0 - w_1 \ell w_1 \cdots - w_p \ell w_p + \text{const.} \qquad (2)$$

where $n = w_1 + w_2 + \cdots + w_p$ and ℓ denotes the natural logarithm. He pointed out that the former formula (1) is a special case of (2) where it is assumed that $f_0 = f_1 = \cdots = f_p$. Ignoring the additive constant the present formula (2) can be rearranged in the form

$$\ell\Omega = -n \sum_{i=0}^{p} g_i \ell\left(\frac{g_i}{f_i}\right) ,$$

where $g_i = w_i/n$. Thus to retain the interpretation that the entropy is the log probability of a distribution we have to adopt, instead of $H(p)$, the quantity

$$B(g;f) = - \sum_i g_i \log\left(\frac{g_i}{f_i}\right)$$

as the definition of the entropy of the secondary distribution g with respect to the primary distribution f. When the distributions g and f are defined with densities $f(x)$ and $g(x)$ the entropy is defined by

$$B(g;f) = - \int g(x) \log\left(\frac{g(x)}{f(x)}\right) dx .$$

When it is necessary to distinguish this quantity from the thermodynamic entropy or the Shannon entropy we will call it the Boltzmann entropy. It is now obvious that $B(g;f)$ provides a natural measure of deviation of g from f.

The equality of the above quantity to the thermodynamic entropy holds only when the former is maximized under the assumption of a given mean energy for an appropriately chosen "primary distribution" f and then multiplied by a proper constant. Thus it can be seen that the Shannon entropy $H(g) = - \Sigma\, g \log g$ obtains the physical meaning of the entropy contemplated by Boltzmann only under very limited circumstances. Obviously $\ell\Omega$ or $B(g;f)$ is the more fundamental concept. This point is reflected in the fact that in Shannon and Weaver (1949) essential use is made not of $H(f)$ but of its derived quantities taking the form of $B(g;f)$.

The Kullback-Leibler (KL) information number is defined by $I(g;f) = -B(g;f)$. Contrary to the formal definition of $I(g;f)$ by Kullback (1959) the present derivation of $B(g;f)$ based on Boltzmann's $\ell\Omega$ clearly explains the difference of the roles played by g and f. The primary distribution f is hypothetical, while the secondary g is factual. Boltzmann (1878) also arrived at a generalization of the exponential family of distributions by maximizing the entropy under certain constraints. These results

demonstrate the fundamental contribution of Boltzmann to the science of statistics. A good summary of mathematical properties of the Boltzmann entropy or the Kullback-Leibler information is given by Csiszar (1975).

6. Entropy and the asymptotic theory of statistics

The Boltzmann entropy appeared, sometimes implicitly, in many basic contributions to statistics, particularly in the area of asymptotic theory. For a pair of distributions $p(\cdot|\theta_1)$ and $p(\cdot|\theta_2)$ from a parametric family $\{p(\cdot|\theta);\ \theta \in \Theta\}$ the deviation of the former from the latter can be measured by $B(\theta_1;\theta_2) = B(p(\cdot|\theta_1);\ p(\cdot|\theta_2))$. This induces a natural topology in the space of parameters.

When θ_1 and θ_2 are k-dimensional parameters given by $\theta_1 = (\theta_{11},\theta_{12},\cdots,\theta_{1k})$ and $\theta_2 = (\theta_{21},\theta_{22},\cdots,\theta_{2k})$, under appropriate regularity conditions we have

$$B(\theta_1;\theta_2) = -\frac{1}{2}(\theta_2 - \theta_1)'E\frac{\partial^2}{\partial\theta'\partial\theta}\log p(x|\theta_1)(\theta_2 - \theta_1) + o(\|\theta_2 - \theta_1\|^2),$$

where $(\partial^2/\partial\theta'\partial\theta)\log p(x|\theta_1)$ denotes the Hessian evaluated at $\theta = \theta_1$ and E the expectation with respect to $p(\cdot|\theta_1)$ and $o(\|\theta_2 - \theta_1\|^2)$ a term of order lower than $\|\theta_2 - \theta_1\|^2 = \sum_i (\theta_{1i} - \theta_{2i})^2$. Obviously $-E(\partial^2/\partial\theta'\partial\theta)\log p(x|\theta_1)$ is the Fisher information matrix. The fact that the Fisher information matrix is just minus twice the Hessian of the entropy clearly shows that it is related to the local property of the topology induced by the entropy.

The likelihood ratio test statistic for testing a specific model, or hypothesis, defined by $\theta = \theta_0$ is given by

$$\lambda_n = \frac{\Pi p(x_i|\theta_0)}{\sup\{\Pi p(x_i|\theta);\ \theta \in \Theta\}},$$

where $(x_1,x_2,\ldots,x_n)$ denotes the sample. If the true distribution is defined by $p(\cdot|\theta)$ we expect that

$$T_n = -\frac{1}{n}\log\lambda_n$$

will stochastically converge to $-B(\theta;\theta_0)$ as n is increased to infinity. The result of Bahadur (1967) shows that under certain regularity conditions it holds that

$$\lim_{n\to\infty}\frac{1}{n}\log P(T_n > t_n|\theta_0) = B(\theta;\theta_0),$$

where t_n denotes the sample value of the test statistic T_n for a particular realization $(x_1, x_2, \ldots, x_n)$. This means that if one calculates the probability of the statistic T_n being larger than t_n, assuming that the data has come from the hypothetical distribution $p(\cdot|\theta_0)$, it will asymptotically be equal to $\exp(nB(\theta;\theta_0))$, where θ denotes the true distribution.

In a practical application the hypothesis will never be exact and the above result says that by calculating the P-value of the log likelihood ratio test we are actually measuring the entropy $nB(\theta;\theta_0)$. This observation eliminates the common misconception that considers the test meaningless due to the certainty of θ_0's being false.

The concept of second order efficiency was introduced by Rao (1961). In that paper he discussed the performance of an estimator obtained by minimizing the Kullback-Leibler information number $\Sigma \pi_r \log(\pi_r/p_r)$, where π_r denotes the probability of the r^{th} cell in a multinomial distribution, defined as a function of a parameter θ, and p_r the observed relative frequency. This estimator can also be characterized as the one that maximizes $B(\pi;p)$, while the maximum likelihood estimate maximizes $B(p;\pi)$.

If we carefully follow the derivation of $B(g;f)$ we can see that the primary distribution f is always hypothetical, while the secondary distribution g is factual. It is interesting to note that Rao has shown that the minimum KL number estimator, defined by the entropy with a factual primary distribution and an hypothetical secondary, is less efficient than the maximum likelihood estimator defined by the more natural definition of the entropy. A similar relation has been observed between the estimators defined by minimizing the chi-square and the modified chi-square that are approximations to $-2B(p;\pi)$ and $-2B(\pi;p)$, respectively. These results suggest that the present interpretation of entropy can produce useful insights not available from the Fisher information which does not discriminate between the primary and secondary distributions.

The relation between the entropy and the asymptotic distribution of the corresponding sample distribution function is discussed by Sanov (1957) and Stone (1974). Some standard references on the relation between the entropy and large sample theory are Chernoff (1956) and Rao (1962).

7. Likelihood, entropy and the predictive point of view

Obviously, one of the most significant contributions to statistics by R. A. Fisher is the development of the method of maximum likelihood. However, there is a definite limitation to the applicability of the idea of maximizing the likelihood.

The limitation can most clearly be seen by the following model selection problem. Consider a set of parametric families $\{p(\cdot|\theta_k)\}$ $(k = 1,2,\ldots,K)$ defined by $\theta_k = (\theta_{k1},\theta_{k2},\ldots,\theta_{kk},\theta_{0k+1},\ldots,\theta_{0K})$. In the k^{th} family, only the first k components of the parameter vector θ_k are allowed to vary but the rest are fixed at some preassigned values $\theta_{0k+1},\ldots,\theta_{0K}$. When data x is given, if we simply maximize the likelihood among the whole families, we always end up with the choice of $p(\cdot|\theta_K^*)$, where θ_K^* denotes the maximum likelihood estimate that maximizes $p(x|\theta_K)$. This means that the method of maximum likelihood always leads to the selection of the most unrestricted parametric model. This is obviously against our expectation. The counseling by a statistician of the choice of the highest possible order whenever fitting a polynomial regression, by the method of maximum likelihood, will certainly lose the trust of his clients.

Fisher was clearly aware of the limitation of his theory of estimation. Pointing out the future possibility of inductive argument which will discuss methods of assigning the functional form of the population by data Fisher (1936) states "At present it is only important to make clear that no such theory has been established". This clearly suggests the necessity of extending the theory of statistical estimation to the situation where several possible parametric models are involved. Such an extension is possible with a proper combination of the predictive point of view and the concept of entropy.

The predictive point of view demanded the generalization of the concept of estimation from that of a parameter to that of the distribution of a future observation. We will call such an estimate a predictive distribution. The basic criterion in this generalized theory of estimation will then be the measure of the goodness of the predictive distribution. The expected deviation of the true distribution from the predictive distribution as measured by the expected entropy EB(true; predictive) will serve for this purpose. Here, the

expectation E is taken with respect to the true distribution of the data used to define the predictive distribution.

In a practical application, except for the data obtained by an artificial sampling scheme, no one knows what is the true distribution. Only through the process of specifying an estimation procedure, or a model, the true distribution obtains a practical meaning. The true distribution may thus be viewed as a conceptual construction that provides a basis for the design of an estimation procedure for a particular type of data. The validity of such a procedure can only be judged by the collective experience of its use by human society. In such a circumstance it becomes important to find objectivity in a statistical inference procedure.

For a parametric model $\{p(\cdot|\theta);\ \theta \in \Theta\}$ we may consider a predictive distribution defined by $f(y|x) = p(y|\theta)$, where y denotes the future observation and x the present data. The goodness of such a predictive distribution is evaluated by the entropy

$$B(*;\theta) = -E_y \log \frac{p(y)}{p(y|\theta)} ,$$

where E_y denotes the expectation with respect to the true distribution denoted by $p(y)$. However, since we have

$$B(*;\theta) = E_y \log p(y|\theta) - E_y \log p(y) ,$$

for the comparative evaluation of θ we may restrict our attention to $E_y \log p(y|\theta)$.

Here we further specify the predictive point of view by assuming that <u>the future observation</u> y <u>is another independent sample taken from the same distribution as that of the present data</u> x. This point of view specifies the objective of our inference. In particular, it specifies that the accuracy of our inference is evaluated only in its relation to the prediction of an observation similar to the present one. This suggests that a too detailed modeling of a situation may not really be necessary for the purpose of inference, thus suggesting the necessity of parsimonious modeling.

One of the important consequences of the present specification of the predictive point of view is that it leads to the observation that the log likelihood $\log p(x|\theta)$ is a natural estimate of $E_y \log p(y|\theta)$. Obviously, by the present predictive point of view, <u>the log likelihood</u> $\log p(x|\theta)$ <u>provides an unbiased estimate of</u> $E_y \log p(y|\theta)$, <u>irrespectively</u>

of the form of the true distribution $p(y)$. This provides a proof of the fact that the log likelihood is an objective measure of the goodness of fit of the distribution $p(\cdot|\theta)$. Thus we see that a definite objectivity is being imparted to statistical inference through the use of log likelihoods. In particular, we can see that the range of the validity of the concept of likelihood is not restricted to one particular parametric family of distributions. This observation constitutes the basis for the solution of the model selection problem considered at the beginning of this section.

8. Model selection and an information criterion (AIC)

We will first show that our basic criterion, the expected entropy, provides a natural extension of the mean squared error criterion. The quality of a predictive distribution $f(y|x)$ is evaluated by the expected negentropy defined by

$$-E_x B(f;f(\cdot|x)) = E_y \log f(y) - E_x E_y \log f(y|x) ,$$

where $f(y)$ denotes the true distribution of y which is assumed to be independent of x and E_x and E_y denote the expectations with respect to the true distributions of x and y, respectively. By Jensen's inequality we have $E_x \log f(y|x) \leq \log E_x f(y|x)$ and we get the additive decomposition

$$-E_x B(f;f(\cdot|x)) = \{E_y \log f(y) - E_y \log E_x f(y|x)\} + \{E_y \log E_x f(y|x) - E_y E_x \log f(y|x)\} .$$

The term inside the first brackets on the right hand side represents the amount of increase of the expected negentropy due to the deviation of $f(y)$ from $E_x f(y|x)$. This term corresponds to the squared bias in the case of ordinary estimation of a parameter. The term inside the second brackets represents the increase of the expected negentropy due to the sampling fuctuation of $f(y|x)$ around $E_x f(y|x)$. This quantity corresponds to the variance. The present result shows why the two different concepts, squared bias and variance, can meaningfully be added.

Having observed that the expected negentropy provides a natural extension of the mean squared error criterion we recognize that the main problem is the estimation of the entropy or the expected log likelihood $E_y \log f(y|x)$ of the predictive distribution. In the case of the ANOVA model discussed by Leonard and Ord the F-test was used for the selection of the model underlying the definition of the final estimate. For the present general model we consider the use of the log likelihood ratio test. The test statistic for the testing of $\{p(\cdot|\theta_k)\}$ against $\{p(\cdot|\theta_K)\}$ is defined by

$$(-2)\{\log p(x|\theta_k^*) - \log p(x|\theta_K^*)\}$$

and is tested as a chi-square with the degrees of freedom $K - k$.

We consider that the test is developed to make a reasonable choice between $p(y|\theta_k^*)$ and $p(y|\theta_K^*)$. From our present point of view this means that the test must be in good correspondence to the choice by $(-2)E_y\{\log p(y|\theta_k^*) - \log p(y|\theta_K^*)\}$. The result of Wald (1943) on the asymptotic behavior of the log likelihood ratio test shows that, when x is a vector of observations of independently identically distributed random variables with the likelihood functions satisfying certain regularity conditions, we have asymptotically

$$E_x[-2\{\log p(x|\theta_k^*) - \log p(x|\theta_K^*)\}] = \|\theta_k^0 - \theta_K^0\|_I^2 + (K - k) ,$$

where E_x denotes the mean of the limiting distribution, $\| \ \|_I$ the Euclidean norm defined by the Fisher information matrix, and θ_k^0 denotes the value of θ_k that maximizes $E_x \log p(x|\theta_k)$, where E_x denotes the expectation with respect to the true distribution $p(x|\theta_K^0)$.

Similarly from the analysis of the asymptotic behavior of the maximum likelihood estimates we have asymptotically

$$E_x[-2E_y\{\log p(y|\theta_k^*) - \log p(y|\theta_K^*)\}] = \|\theta_k^0 - \theta_K^0\|_I^2 - (K - k) ,$$

where the restricted predictive point of view is adopted and x and y are assumed to be independently identically distributed.

From these two results it can be seen that as a measurement of $(-2)E_y\{\log p(y|\theta_k^*) - \log p(y|\theta_K^*)\}$ the log likelihood ratio test statistic $(-2)\{\log p(x|\theta_k^*) - \log p(x|\theta_K^*)\}$ shows an upward bias by the amount of $2(K - k)$. If we correct for this bias then we get $\{-2 \log p(x|\theta_k^*) + 2k\} - \{-2 \log p(x|\theta_K^*) + 2K\}$ as a measurement of the difference of the entropies of the models specified by $p(\cdot|\theta_k^*)$ and $p(\cdot|\theta_K^*)$. This observation leads to the conclusion that the statistic $-2 \log p(x|\theta_k^*) + 2k$ should be used as a measure of the badness of the model specified by $p(\cdot|\theta_k^*)$ (Akaike, 1973). The pseudonym AIC adopted by Akaike (1974) for this statistic is the abbreviation of "an information criterion" and is symbolically defined by

$$\text{AIC} = -2 \log(\text{maximum likelihood}) + 2 (\text{number of parameters}) ,$$

where log denotes natural logarithm.

If the log likelihood ratio test is considered as a measurement of the entropy difference then the above observation suggests that from our present point of view we should choose the model with smaller value of AIC. If we follow this idea we get an estimation procedure which simultaneously realizes the model selection and parameter estimation. An estimate thus obtained is called a minimum AIC estimate (MAICE). Now it is a simple matter to see that the critical level 2 of the F test by Leonard and Ord corresponds to the factor 2 of the second term in the definition of AIC.

One important observation about AIC is that it is defined without specific reference to the true model $p(\cdot|\theta_K^0)$. Thus, for any finite number of parametric models, we may always consider an extended model that will play the role of $p(\cdot|\theta_K)$. This suggests that AIC can be useful, at least in principle, for the comparison of models which are non-nested, i.e., the situation where conventional log likelihood ratio test is not applicable.

We will demonstrate the practical utility of AIC by its application to the multidimensional contingency table analysis discussed by Goodman (1971). Observing the frequency $f_{ijk\ell}$ in the cell (i,j,k,ℓ) of a 4-way contingency table $(i = 1,2,\ldots,I;\ j = 1,2,\ldots,J;\ k = 1,2,\ldots,K;\ \ell = 1,2,\ldots,L)$ with $\Sigma_{ijk\ell} f_{ijk\ell} = n$ the basic model is specified by the parametrization

$$\log F_{ijk\ell} = \theta + \lambda_i^A + \cdots + \lambda_\ell^D + \lambda_{ij}^{AB} + \cdots + \lambda_{k\ell}^{CD} + \lambda_{ijk}^{ABC} + \cdots + \lambda_{jk\ell}^{BCD} + \lambda_{ijk\ell}^{ABCD},$$

where $F_{ijk\ell}$ denotes the expected frequency and the λ's satisfy the condition that any sum with respect to one of the suffices is equal to zero. The characters A, B, C, D symbolically denotes the group of parameters that are related with the factors denoted by these characters. Hypotheses are defined by putting some of the parameters equal to zero.

Goodman discussed the application to the analysis of detergent user data which included information on the following four factors: the softness of the water used (S), the previous use of a brand (U), the temperature of the water used (T) and the preference of a brand over the other (P). In the following Table 1 the initial portion of Goodman's Table 3 is shown with the corresponding AIC's. In the Goodman's modeling when a higher order effect is considered all the corresponding lower order effects are included in the model.

Table 1: Goodman's analysis of consumer data

Hypothesis	Estimated Group of Parameters	Degrees of Freedom	(-2)Log Likelihood Ratio	AIC*
1	None	23	118.63	72.63
2	S, P, T, U	18	42.93	6.93
3	All the pairs	9	9.85	- 8.15
4	All the triplets	2	0.74	- 3.26
5	PU, S, T	17	22.35	-11.65
6	PU, S	18	95.56	59.56
7	PU, T	19	22.85	-15.15
8	PU, PT	18	18.49	-17.51
9	PT, U	19	39.07	1.07
10	PU, PT, ST	14	11.89	-16.11
11	PU, PT, S	16	17.99	-14.01

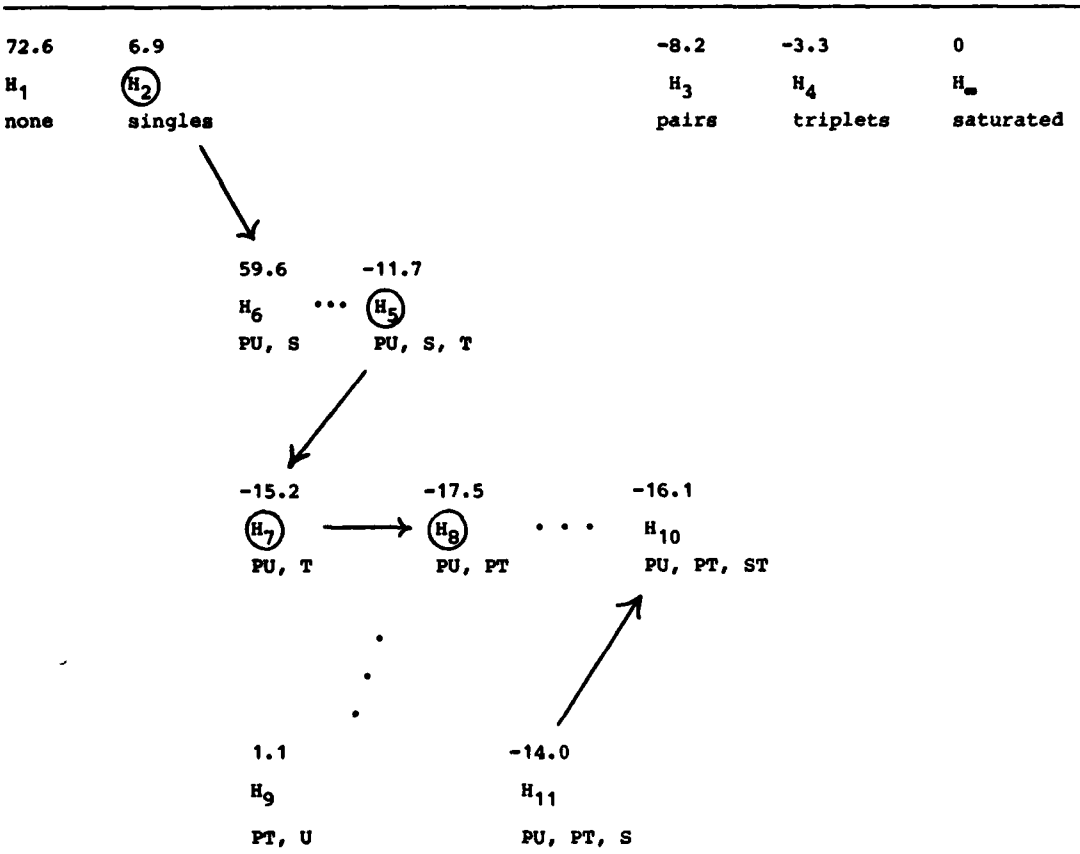
*AIC = (-2)(Log Likelihood Ratio) - 2(Degrees of Freedom)
= AIC(i) - AIC(∞), where AIC(i) denotes the original AIC of H_i and AIC(∞) denotes that of the saturated model with all the parameters unrestricted.

Goodman asserts that H_1 and H_2 do not fit the data but H_3 and H_4 do, where H_i denotes hypothesis numbered by i. By the present definition of AIC the negative signs of AIC for H_3 and H_4 means that the corresponding models are preferred to the saturated non-restricted model. This corresponds to Goodman's assertion. The AIC already suggests that H_4 is an over-fit and Goodman actually proceeds to the detailed analysis of H_3 and arrives at H_5.

The significances of S and T are then respectively checked by comparing H_6 and H_7 with H_5. The hypothesis H_8 is then judged to be an improvement over H_7. The

effect of PU is then confirmed by comparing H_8 with H_9. Further elaboration of H_8 leads to H_{10}. However, its improvement over H_8 is not considered to be significant, although the effect ST is judged to be significant by the comparison of H_{10} with H_{11}. The path of Goodman's stepwise search is schematically represented by Table 2.

Table 2. The path of Goodman's stepwise search and the corresponding AIC's*



*The number above each hypothesis denotes the AIC relative to that of H_∞.

Table 2 shows that we come to one and the same conclusions as those obtained by Goodman with the choice of 5% as the critical level, simply by choosing models with lower

9. Entropy maximization principle and the Bayes procedure

The discussion of the concept of true model an its relation to entropy clearly shows that there is no end in statistical model building. All we can do is to produce better models. When we admit this then it is easy to accept the following very modest, yet very productive, view of statistics; all statistical activities are directed to maximize the expected entropy of the predictive distribution in each particular application. We call this the entropy maximization principle (Akaike, 1977). The minimum AIC procedure may be considered as a realization of this principle. The generality of this principle can be seen by the following discussion of Bayesian modeling.

Consider the set of models given by $\{g_k(\cdot);\ k = 1,2,\ldots,K\}$, where $g_k(y)$ denotes a predictive distribution specified by the parameter k. Assume that we consider the use of a random mechanism for the selection of the predictive distribution. Our preference of the models is represented by the distribution of probabilities $w_k(x)$ of selecting the k^{th} model, where $w_k(x)$ is specified by combining our knowledge of the problem and the data x. However, irrespectively of the form of the true distribution of y, the following relation holds

$$E_y \log\Big\{\sum_{k=1}^{K} g_k(y)w_k(x)\Big\} \geq \sum_{k=1}^{K} w_k(x)E_y \log g_k(y) ,$$

where E_y denotes the expectation with respect to the true distribution of y. This means that the entropy of the true distribution with respect to the averaged distribution $\Sigma\, g_k(y)w_k(x)$ is always greater than or equal to that with respect to the distribution chosen by the random mechanism. The entropy maximization principle suggests that we should consider the use of the averaged distribution $\Sigma\, g_k(y)w_k(x)$ as our predictive distribution rather than the distribution to be chosen by the random mechanism. Taking into account the fact that conventional model selection procedure is realized by a particular choice of $w_k(x)$ which takes either the value 0 or 1, the present result suggests the possibility of improved modeling by extending the basic set of models from

$$\{g_k(\cdot);\ k = 1,2,\ldots,K\} \quad \text{to} \quad \{\Sigma_k g_k(\cdot)w_k;\ w_k \geq 0,\ \Sigma_k w_k = 1\} .$$

values of AIC. The fact that AIC does not require the table look-up of the chi-squares with different degrees of freedom adds to the significance of this result. Since AIC is defined with a unique scaling unit it allows easy extraction of useful information from a collection of fitted models. For example, by comparing the difference of AIC's of H_7 and H_5 with that of H_8 and H_{11}, we can clearly see the deteriorating effect of including S into the model. Also the direct comparison of H_6 and H_7, not possible by the log likelihood ratio test, is now possible by AIC and the inferiority of H_6 that contains S is clearly recognizable. The ability of AIC to allow the researcher to extract global information from the result of fitting a large number of models is a unique characteristic that is not shared by the conventional model selection procedure realized by some ad hoc application of significance tests.

AIC attracted much attention from people in both theoretical and applied fields of statistics. In particular the 1974 paper (Akaike, 1974) has been spotted by the Institute for Scientific Information as one of the most frequently cited papers in the area of engineering, technology and applied sciences, with the frequency of citations over 180 during 1974-81 (Akaike, 1981). Some of the theoretical works related with AIC are the discussion of the asymptotic equivalence of the minimum AIC procedure to cross-validation by M. Stone (1977b), modifications of the criterion by Schwarz (1978) and Hannan and Quinn (1979), discussions of the relation to the Bayes procedure by Zellner (1978), Atkinson (1980) and Smith and Spiegelhalter (1980) and discussions of the optimality of the MAICE type procedure by Akaike (1978a), Shibata (1980) and C. J. Stone (1982). The inherent relation between the magic number 2 and the predictive point of view can be seen also by the works by Geisser and Eddy (1979) and Leonard (1977).

When the number of possible alternatives is increased the MAICE procedure may tend to be sensitive to sampling fluctuations. One solution to this problem is to use some averaging procedure as is discussed in Akaike (1979). However, this brings us closer to Bayesian modeling which is going to be discussed in the next section.

The problem now is how to define $w_k(x)$. Since the distribution $w_k(x)$, which we will call the inferential distribution, is introduced to define a predictive distribution, we will consider the more general problem of the selection of a predictive distribution. Assume that the variable x takes a finite number of discrete values $x = 1,2,\ldots,I$. Before the observation of x we consider the selection of the predictive distribution of x. We assume that the possible predictive distributions of x are also parametrized as $f_k(x)$. Since x is not available yet we consider the use of a probability distribution w_k over k, defined independently of x. Thus we are specifying a probability distribution $w_k f_k(x)$ over (k,x).

When the observation produces $x = x_0$ a Bayesian will say that we should follow the Bayes procedure and replace the distribution $w_k f_k(x)$ by the distribution $w(k,x)$ which is defined by

$$w(k,x) = \frac{w_k f_k(x_0)}{\sum_k w_k f_k(x_0)} \quad \text{for} \quad x = x_0 ,$$

$$0 \qquad \text{otherwise} .$$

However, this suggestion is not based on any clearly defined principle. One could have chosen any $w(k,x)$ as a function of x_0, if only the distribution is limited to $x = x_0$.

There is an essential analogy between the Boltzmann's derivation of the exponential family of distributions for energy and the Bayes procedure. To see this we consider more generally an arbitrary distribution $\pi(k,x)$ over (k,x) and try to find a distribution $w(k,x)$ concentrated on $\{(k,x_0)\}$ and such that the Boltzman entropy with respect to the original $\pi(k,x)$ is maximum. This leads to the maximization of

$$\sum_x \sum_k w(k,x)\{\log \pi(k,x) - \log w(k,x)\} + \lambda\Big(\sum_k w(k,x_0) - 1\Big) ,$$

where λ is the Lagrange multiplier. The solution is given by

$$w(k,x) = \frac{\pi(k,x_0)}{\sum_k \pi(k,x_0)} \quad \text{for} \quad x = x_0 \ ,$$

$$0 \qquad \text{otherwise} \ .$$

This result characterizes the transition from the original distribution to the conditional distribution as the most conservative action that conforms to the observation of the data x_0 yet otherwise maximally retains the structure of the originally assumed distribution. We will call this particular application of the maximum entropy method of probability distribution generation by the name of conditioning principle.

Coming back to Bayesian modeling we can now see that the assumption of the original distribution $\pi(k,x)$ and the conditioning principle leads to the use of the "posterior distribution" $w(k,x)$ as the inferential distribution $w_k(x)$. That such a definition of the inferential distribution is a reasonable one can be shown as follows. First we assume that when k is given y and x are independent and the distribution is given by $g_k(y)f_k(x)$. The expected performance of a predictive distribution $h(y|x)$ is then evaluated by $E_k E_{x|k} E_{y|k} \log h(y|x)$, where E_k denotes the expectation with respect to the distribution w_k and $E_{x|k}$ and $E_{y|k}$ denote the expectations with respect to $f_k(x)$ and $g_k(y)$, respectively. We have

$$E_k E_{x|k} E_{y|k} \log h(y|x) = \sum_x f(x) \sum_y \sum_k g_k(y) w(k|x) \log h(y|x) \ ,$$

where $f(x) = \Sigma\, f_k(x) w_k$ and $w(k|x) = f_k(x)w_k/f(x)$. This quantity is maximized by putting

$$h(y|x) = \sum_k g_k(y) w(k|x) \ ,$$

which means that, as long as we assume the validity of the original probabilistic set-up, the use of the posterior distribution $w(k|x)$ as the inferential distribution is the best choice. This result is recognized earlier by Kerridge (1961) and Aitchison (1975).

10. Statistical inference and Bayesian modeling

What the result of the preceding section has shown is that the conditioning principle leads to the best choice of the inferential distribution under the assumption of the validity of the Bayesian model defined by $f_k(y)f_k(x)w_k$. What would happen when we are uncertain about the choice of the "prior distribution" w_k?

Here we recall our basic observation that statistical model building is an unending process. This means that the validity of a model can only be established by a careful analysis of other possibilities. This leads to the situation where we have several alternative prior distributions $w_k^{(i)}$ $(i = 1,2,\ldots,I)$. Here we have to assume a (hyper) prior distribution $\pi(i)$ over i's. When the data x is observed the posterior probability $p(i|x)$ of the i^{th} model is given by the relation

$$p(i|x) = f^{(i)}(x)\pi(i) ,$$

where $f^{(i)}(x)$ is the likelihood of the i^{th} Bayesian model defined by

$$f^{(i)}(x) = \sum_k f_k(x)w_k^{(i)} .$$

Thus, even when we do not know how to specify $\pi(i)$, we can see how much relative support was given to each model by the observation x. Good (1965) called the procedure of hyperparameter estimation by maximizing the likelihood of a Bayesian model the type II maximum likelihood procedure. The use of the likelihood for the assessment of a Bayesian model is demonstrted in an illuminating paper by Box (1980). The application to the very practical problem of seasonal adjustment is discussesd by the present author (Akaike, 1980a).

The discussion of Bayesian modeling will never be complete unless we provide a procedure for the modeling of the situation where no further prior information is available for the modeling. The concept of entropy again finds an interesting application in this type of situation. It has been shown tht the well-known Jeffreys' ignorance prior distribution (Jeffreys, 1946) can be given an interpretation as the locally or globally impartial prior distribution (Akaike, 1978b).

However, this concept is essentially dependent on the continuity of the parameter involved. Recently the present author applied the predictive point of view and the concept of entropy to define a prior distribution that "lets the data speak most". For the Bayesian model discussed in the preceding section this prior distribution, called the minimum information prior distribution, is defined as the one that maximizes

$$I(w) = \sum_{y} \sum_{x} h(y,x) \log \frac{h(y,x)}{g(y)f(x)} ,$$

where $g(y) = \Sigma\, g_k(y)w_k$, $f(x) = \Sigma\, f_k(x)w_k$ and $h(y,x) = \Sigma\, g_k(y)f_k(y)w_k$. The strict predictive point of view demands us to put $g_k(y) = f_k(y)$. It has been observed that this definition leads to interesting non-trivial specifications of the prior distribution over a finite discrete set of alternatives (Akaike, 1982).

Related works in this area are those by Zellner (1977) and Bernardo (1979) based on the earlier work of Lindley (1956) who discussed the use of the Shannon entropy in statistics.

Do these formal procedures of generating prior distributions produce useful results? The anwer can be obtained only through the detailed analysis of the final output of each Bayesian model thus obtained. An example of such an analysis is given by Akaike (1980b) where admissibility is proved for the James-Stein type estimator of a multivariate normal distribution obtained by applying the ignorance prior to the hyperparameter of a prior distribution.

Here again we are reminded of the attitude of Boltzmann who considered the justification of the primary distribution used in the derivation of the distribution of the energy could only be obtained through the observation of the validity of the final result. The use of a Bayesian procedure can only be justified when the procedure produces good results for those data which are "similar" to the present one and for which unequivocal judgment of the results is possible.

11. Conclusion

It is now clear that the predictive point of view, particularly in its strict form, and the concept of entropy can produce a unifying view of statistics. This view is not only conceptually simple and unifying but also practically very productive. The notorious difficulty of the significance test of multiple hypotheses is given a practical solution by AIC. The historical split between the Bayesian and non-Bayesian is now eliminated.

The entropy maximization principle which is obtained by combining the predictive point of view with the concept of entropy clearly states that the search for better models is the purpose of statistical data analysis. Bayesian modeling will often be useful in improving the presently existing non-Bayesian models. However, models are formulations of our past experiences and only new interesting real problems can stimulate the development of useful models. The fundamental contribution by Boltzmann came from the deep study of one particular real problem. Thus we can see that for the development of statistics the main emphasis should be placed on the search for important practical problems. This forms the conclusion of the present paper.

Acknowledgements

The author is grateful to A. P. Dawid and T. Leonard for helpful comments.

REFERENCES

Aitchison, J. (1975) Goodness of prediction fit. Biometrika 62, 547-554.

Akaike, H. (1969) Fitting autoregressive models for prediction. Ann. Inst. Statist. Math. 21, 243-247.

Akaike, H. (1970) Statistical predictor identification. Ann. Inst. Statist. Math. 22, 203-217.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. (B. N. Petrov and F. Csaki, eds.) Second International Symposium on Information Theory, Budapest, Akademiai Kiado, pp. 267-281.

Akaike, H. (1974) A new look at the statistical model identification. IEEE Trans. Automat. Contr. AC-19, 716-723.

Akaike, H. (1977) On entropy maximization principle. (P. R. Krishnaiah, ed.) Applications of Statistics, Amsterdam, North-Holland, pp. 27-41.

Akaike, H. (1978a) A Bayesian analysis of the minimum AIC procedure. Ann. Inst. Statist. Math. 30, A, 9-14.

Akaike, H. (1978b) A new look at the Bayes procedure. Biometrika 65, 53-59.

Akaike, H. (1979) A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. Biometrika 66, 53-59.

Akaike, H. (1980a) Seasonal adjustment by a Bayesian modeling. J. Time Series Analysis 1, 1-13.

Akaike, H. (1980b) Ignorance prior distribution of a hyperparameter and Stein's estimator. Ann. Inst. Statist. Math. 33A, 171-179.

Akaike, H. (1981) Abstract and commentary on "A new look at the statistical model identification". Current Contents, Engineering, Technology and Applied Sciences, 12, No. 51, 22.

Akaike, H. (1982) On minimum information prior distributions. Technical Summary Report No., Mathematics Research Center, University of Wisconsin-Madison,

Atkinson, A. C. (1980) A note on the generalized information criterion for choice of a model. Biometrika 67, 413-418.

Bahadur, R. R. (1967) An optimal property of the likelihood ratio statistic. (L. M. LeCam and J. Neyman, eds.) Proc. 5th Berkeley Symp. Math. Statist. and Prob. 1, Berkeley, University of California Press, pp. 13-26.

Bernardo, J. M. (1979) Reference posterior distributions for Bayesian inference, (with discussion). J. Roy. Statist. Soc. B41, 113-147.

Boltzmann, L. (1872) Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. Wiener Berichte 66, 275-370.

Boltzmann, L. (1877a) Bemerkungen über einige Probleme der mechanischen Wärmetheorie. Wiener Berichte 75, 62-100.

Boltzmann, 1. (1877b) Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respective den Sätzen über das Wärmegleichgewicht. Weiner Berichte 76, 373-435.

Boltzmann, L. (1878) Weitere Bemerkungen über einige Plobleme der mechanischen Wärmetheorie. Wiener Berichte 78, 7-46.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. J. R. Statist. Soc. A143, 383-430.

Chernoff, H. (1956) Large sample theory - Parametric case. Ann. Math. Statist. 27, 1-22.

Csiszar, I. (1975) I-divergence geometry of probability distributions and minimization problems. Ann. Prob. 3, 146-158.

Fisher, R. A. (1935) The fiducial argument in statistical inference. Annals of Eugenics 6, 391-398. Paper 25 in Contributions to Mathematical Statistics (1950), New York, Wiley.

Geisser, S. and Eddy, W. F. (1979) A predictive approach to model selection, J. Amer. Statist. Assoc. 74, 153-160.

Good, I. J. (1965) The Estimation of Probabilities. Cambridge, Massachusetts, M. I. T. Press.

Goodman, L. A. (1971) The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics 13, 33-61.

Guttman, I. (1967) The use of the concept of a future observation in goodness-of-fit problems. J. R. Statist. Soc. B29, 83-100.

Hannan, E. j. and Quinn, B. G. (1979) The determination of the order of an autoregression. J. R. Statist. Soc. B41, 190-195.

Jaynes, E. T. (1957) Information theory and statistical mechanics. Phys. Rev. 106, 620-630 and 108, 171-182.

Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems. Proc. R. Soc. London A186, 453-461.

Kerridge, D. F. (1961) Inaccuracy and inference. J. R. Statist. Soc. B23, 184-194.

Kullback, S. (1959) Information Theory and Statistics. New York, Wiley.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. Ann. Math. Statist. 22, 79-86.

Leonard, T. and Ord, K. (1976) An investigation of the F-test procedure as an estimation short-cut. J. R. Statist. Soc. B38, 95-98.

Leonard, T. (1977) A Bayesian approach to some multinomial estimation and pre-testing problems. J. Amer. Statist. Assoc. 72, 869-876.

Lindley, D. V. (1956) On a measure of the information provided by an experiment. Ann. Math. Statist. 27, 986-1005.

Mallows, C. L. (1973) Some comments on C_p. Technometrics 15, 661-675.

Pearson, K. (1929) Laplace, being extracts from lectures delivered by Karl Pearson. Biometrika 21, 202-216.

Rao, C. R. (1961) Asymptotic efficiency and limiting information. (J. Neyman, ed.) Proc. 4th Berkeley Symp. Math. Statist. and Prob. 1, Berkeley, University of California Press, pp. 531-548.

Rao, C. R. (1962) Efficient estimates and optimum inference procedures in large samples. J. Roy. Statist. Soc. B24, 46-72.

Sanov, I. N. (1957) On the probability of large deviations of random variables (in Russian). Mat. Sbornik N. S. 42 (84), 11-44. (English translation in Selected Transl. Math. Statist. Prob. 1 (1961), 213-244.)

Schwarz, G. (1978) Estimating the dimension of a model. Ann. Statist. 6, 461-464.

Shannon, C. E. and Weaver, W. (1949) The Mathematical Theory of Communication. Urbana, University of Illinois Press.

Shibata, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameter of a linear process. Ann. Statist. 8, 147-164.

Smith, A. F. M. and Spiegelhalter, D. J. (1980) Bayes factors and choice criteria for linear models. J. R. Statist. Soc. B42, 213-220.

Stigler, S. M. (1975) The transition from point to distribution estimation. ISI Bulletin, 40th ISI Meeting, vol. 2, 332-340.

Stone, C. J. (1982) Local asymptotic admissibility of a generalization of Akaike's model selection rule. Ann. Inst. Statist. Math. 34A, 123-133.

Stone, M. (1974) Large deviations of empirical probability measures. Ann. Statist. 2, 362-366.

Stone, M. (1977a) Asymptotics for and against cross-validation. Biometrika 64, 29-35.

Stone, M. (1977b) Asymptotic equivalence of choice of models by cross-validation and Akaike's criterion, J. R. Statist. Soc. B39, 44-47.

Wald, A. (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Tran. Amer. Math. Soc. 54, 426-482.

Zellner, A. (1977) Maximal data information prior distributions. (A. Aykac and C. Brumat, eds.) New Developments in the Applications of Bayesian Methods, Amsterdam, North-Holland, 211-232.

Zellner, A. (1978) Jeffreys-Bayes posterior odds ratio and the Akaike information criterion for discriminating between models. Economic Letters 1, 337-342.

HA:scr

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 2397	2. GOVT ACCESSION NO.: AD-A120 956
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): PREDICTION AND ENTROPY	5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Hirotugu Akaike	8. CONTRACT OR GRANT NUMBER(s): DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709	12. REPORT DATE: June 1982 13. NUMBER OF PAGES: 33
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Entropy
Predictive distribution
Likelihood
Model selection
Bayes procedure
Information
AIC
Entropy maximization principle

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The emergence of the magic number 2 in recent statistical literature is explained by adopting the predictive point of view of statistics with entropy as the basic criterion of the goodness of a fitted model. The historical development of the concept of entropy is reviewed and its relation to statistics is explained by examples. The importance of the entropy maximization principle as the basis of the unification of conventional and Bayesian statistics is discussed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

1-83

DTIC